# Things disappear* from the internet.

* sometimes, are deliberately disappeared

## Care about stuff on the web?

art
zines
fan fic
videos/tiktoks
photo galleries
social media hashtags
departed loved ones' posts
museum & archival materials
 that we have always existed
what really happened
cultural memory


human rights
resistance
anger
joy

## *You* need to act to archive it.

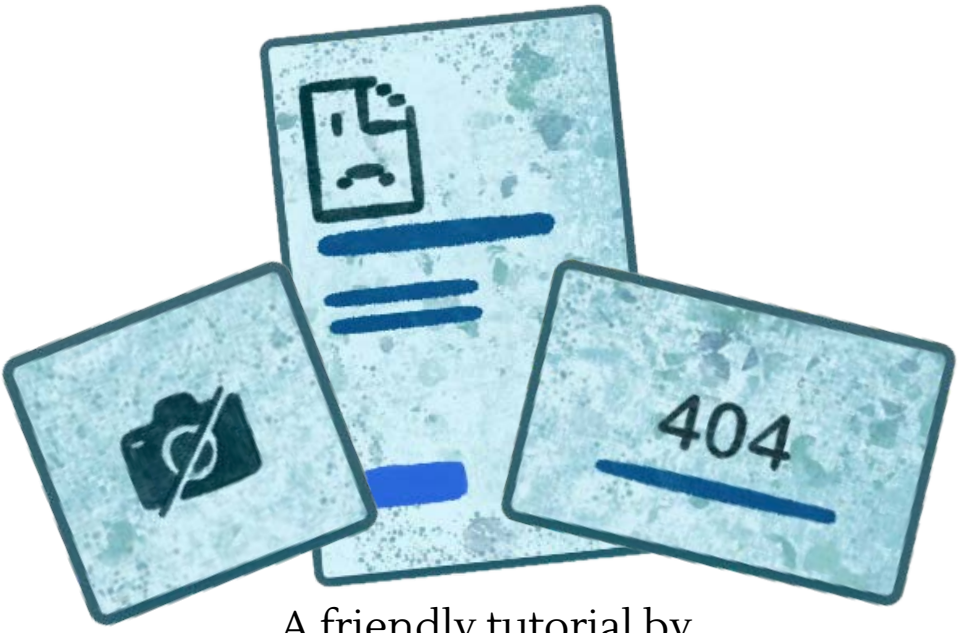
## Anyone can do it— this zine shows you how.

***DIY Web Archiving Zine***

*Dombrowski, Kijas, Kreymer, Walsh, Visconti*

# DIY Web Archiving

FOR THE WEB (& WORLD) **YOU CARE ABOUT**




A friendly tutorial by
Tessa Walsh, Ilya Kreymer (Webrecorder);
Quinn Dombrowski, Anna Kijas (SUCHO);
& Amanda Wyatt Visconti

(CC BY-NC: see last page for how to cite this zine,
but tl;dr please DO share digital+printed copies freely!)

# DIY Web Archiving




*Reverse photo: double-faced lion, SUCHO copy of National Folk Decorative Art Museum item; gallery.sucho.org/items/show/25. Spread graphics: Webrecorder.net.*

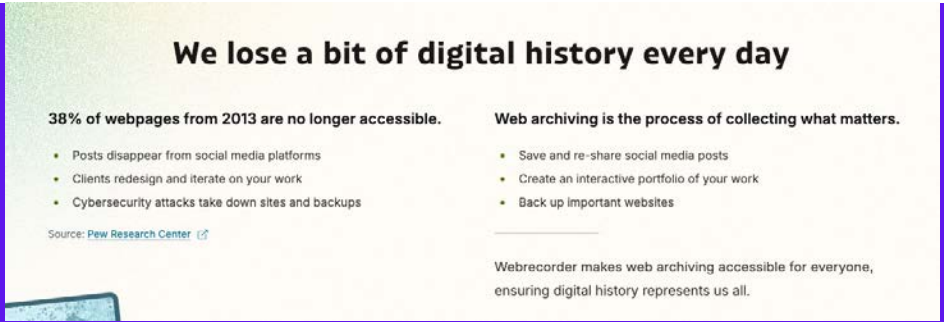
# Why DIY web archiving?

YOU can be part of making sure the stuff you care about on the web doesn't disappear! Anything you access via the web, including datasets, fanfic, encyclopedia articles, videos, websites, online museum exhibits, online community forums, news articles, and more.

We're hoping to give you concrete tools and steps that you can take like right now today to start preserving things that you are concerned might no longer be there in a month, 3 months, a year. This zine introduces you to accessible tools and best practices to get started, and why YOU should act rather than assuming others will take care of things.

Don't overthink it! It's tempting to try to coordinate this work, create giant lists to make sure we don't have too much duplication. We advise you don't get distracted building that kind of infrastructure when there are things needing archiving right now that might not always be there.

It's okay if we end up with a tapestry of partially overlapping archived copies, because everyone everyone captured as many things as we could that we think are at risk.



# Why not let someone else archive it?

Why do DIY web archiving at all, rather then let other services or people handle things? Especially when there are options like the Internet Archive's Wayback Machine that you can just send URLs you want saved, not think further about it, and feel reasonably confident that they'll go make a copy and then you can go retrieve that copy later.

For federal government data, there's a group that's doing end-of-term archiving where anyone now can submit a URL and then they'll go crawl that as part of their end of term web archiving crawl.


It's important that people are doing this! But it's risky for us to rely on any one, or 2 or even 3 services or groups to capture and store things you care about (e.g. the Internet Archive went temporarily offline in 2024 due to a cyber attack 😬). If you care about it, you should take steps to ensure that it continues to exist.

# Anyone can do this!

Downloading a copy yourself and storing it someplace that you control is really the best way to do it. It does take work and care to do. This stuff doesn't just magically happen. But at the same time, it also doesn't require being a super technical person.

As part of SUCHO (Saving Ukranian Cultural Heritage Online initiative), we taught everyone from literal kindergartners, to retirees unfamiliar with computing, how to do web archiving. Your kids, your parents can do this. This is a way to actively contribute to protecting the things that matter to you.

Given the state of the world right now, it's worth considering what kinds of websites can be reached with a slow internet connection, & non-digital methods for copying and distributing things that you care about (like the paper version of this zine!). Things people can leave lying around for others to find, hand out—that don't require engaging with online social media platforms, or even having reliable and unmonitored internet. *(If designing for these needs intrigues you, search the web for "digital humanities minimal computing", aka "mincomp")*

*Illustration: Webrecorder.net graphics.*

# Step 1: Making copies

OUR GOAL IS BUILDING WEB ARCHIVING TOOLS FOR ALL.

ILYA KREYMER, WEBRECORDER PROJECT CREATOR

"LOCKSS" (LOTS OF COPIES KEEPS STUFF SAFE*) GETS TO THE HEART OF THE MATTER... OUR ETHOS IS THAT WEB ARCHIVING SHOULD BE OPEN & EASY FOR EVERYONE.

TESSA WALSH, ARCHIVIST & WEBRECORDER SENIOR DEVELOPER


=> **webrecorder.net**

We'll show you some tools we like, so you can get to work protecting online work you care about!

Webrecorder's lowest-barrier and easiest-to-use tool is the Archiveweb.page browser extension, which lets you make copies of webpages and websites.

For working with larger websites, we'll then cover Webrecorder's Browsertrix tool. Because there are extensive online Webrecorder tutorials, this zine will give a skimmable overview and direct you to those docs for deeper info.

All Webrecorder's tools (Webrecorder.net) are free and open-source (i.e. the code, aka source, that runs these tools is posted publicly so that anyone can read it and potentially build on it).

** LOCKSS is the name of a Stanford digital archiving initiative, lockss.org*

# Webrecorder's browser plugin

Archiveweb.page works by recording your traffic as you browse. It writes that traffic to standard web archiving file formats, which can be replayed offline later, or hosted on a different site as an embedded browsable archive.

Unlike a screen recording, a viewer of the archive can actually interact with the captured pages, e.g. following different paths around the pages, rather than only the path you followed while you were archiving.

There are 2 options for using Webrecorder's Archiveweb.page tool, both with the same features:

1. Visit Archiveweb.page to install it as a browser plugin you can easily use while browsing. You'll need to use the Chrome browser or another Chromium-based browser (e.g. Brave) for this (can download it for free just for your archiving work, if you don't already use).

1. Or, you can install the desktop app version from Archiveweb.page (using any browser!)

ARCHIVING: Once you've got a site you want to record...

1. Create a collection: click the browser extension icon...
    a. A popup appears; click its dropdown under "save to:", then select "+ new archiving session"
    b. Type in a name that'll help you remember what site you were recording, and click the check icon button.

2. Initiate archiving: click the “start archiving” button.
    - The webpage will reload, with a banner stating “’Webrecorder ArchiveWeb.page’ started debugging this browser” appearing at the top of the page (this is fine, just related to how the tool is coded).

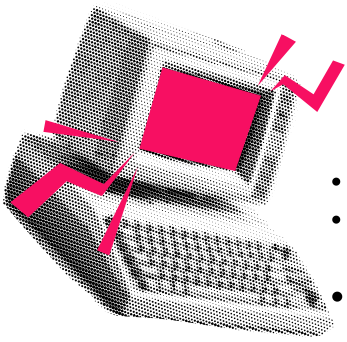
3. Capture session: it’s time to archive! As we browse page by page, each page that we visit is getting recorded into our archive.

- Click around the page.
- Click on assets you want archived (e.g. for an embedded YouTube video, just make sure the video is loaded; you don’t need to watch the whole thing).
- Click on hyperlinks.

4. Check the browser extension—if it’s yellow and says that URLs are pending, that means it's still doing work capturing stuff on the pages. Don’t click “stop” until these have finished.

- You can also see the size of the archive and number of pages increase as you browse from the same menu.

5. To stop recording, click “stop” in the extension icon menu.



Every website is weird in its own way! If you run into problems trying to archive a site, visit Webrecorder’s pages for:

- full tutorial: archiveweb.page/guide
- troubleshooting problems: archiveweb.page/en/troubleshooting
- contact/help archiveweb.page/en/contact

VIEWING ARCHIVES ONLINE: Once created, web archives remain in the browser and can be accessed at any time from the extension home page (house icon in extension menu; or from the Archiveweb.page site, clicking on “Browse Web Archives stored in this browser”). The archive opens in our browser-based web archive replay system, Replay Webpage (ReplayWeb.Page). It’s all being loaded from your archive, not the live internet.

You can organize your archives into multiple distinct collections, which can be browsed and accessed separately from other archives. You can also search individual pages by URL or the text that’s found on a webpage.

DOWNLOADING ARCHIVES: You can download your entire collection, or specific pages from the collection; as a WARC file, or the recommended WACZ format* (which makes sharing and transferring web archives easier, and includes both the WARC files and additional metadata).

Using the WACZ format will allow your archives to load quickly using ReplayWeb.page. ReplayWeb.page is entirely in your browser: when you upload files there to explore them, you're not uploading them to a place where Webrecorder can see them. You’re just loading locally in your browser; the whole Replay system is client-side (happening on your computer, not being sent to and processed on someone else’s computer/server/cloud). This is good for retaining control and privacy of your archive!


** A file format recommended by the Library of Congress, among others.*

DOWNLOADING ALL OR PART OF YOUR ARCHIVE:

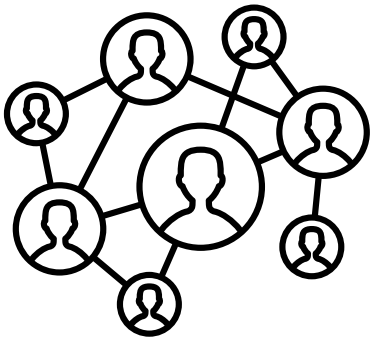
- To download an ENTIRE COLLECTION, click the download icon under “Actions”.
- To download SPECIFIC PAGES from your collection, select the collection, check off the web pages you want to download, choose how you want it downloaded (choose “selected”; and WARC vs. WACZ format—the latter is recommended). Click the download icon.

When we open up that WACZ zip file, we can see that single zip conveniently holds everything needed to replay the archive: an archive directory with the files of our web archiving traffic from when we ran the tool while browsing the pages, and useful features like indexes and page lists.

---

This tool can archive webpages that are behind paywalls, as well as algorithmically generated content like social media feeds that differ among users and page visits as with Bluesky.

For the latter, visiting settings to enable the “archive local storage” option for it to work. As such site archives include things private to your browser (e.g. login credentials), you’ll likely want to store them in a collection that you keep private to yourself rather than download to share with others.



# Webrecorder's Browsertrix

If you want to archive more than what you can see when clicking around the web with the Archiveweb.page extension running, like entire websites or a whole section of a web domain, you'll want to use Webrecorder's Browsertrix tool: a web archiving platform combining a bunch of Webrecorder tools into one place in your browser.

Browsertrix is offered by Webrecorder as a hosted subscription service, or if you're technically inclined you can host it yourself on your own infrastructure.

1. Sign up for a Browsertrix account (docs.browsertrix.com/user-guide/signup) and log in.
2. You'll get an email asking you to name your organization (can be "my personal archiving" or the name of a team you're archiving with/for).
3. Set up a "crawl workflow" to tell Browsertrix what to archive for you.

- The easiest option is the 1st one, "pages in the same directory"; for a site like sucho.org where all pages are contained inside that top-level link, you'll see the tool show you that's everything that will be captured.
- (For more thorough tutorials on these settings, and for troubleshooting, see docs.browsertrix.com.)
- Review your settings, save, and run the crawl.

ONCE YOU'VE STARTED THE CRAWL, you'll see both data about how far along the crawl is, as well as the actual browser window where you watch the crawl in real time if you want, page by page, and a list of upcoming pages the crawl will get to next. If you see pages you don't want to crawl, click the "edit exclusions" button to exclude those pages from the crawl.

ONCE THE CRAWLS SHOWS AS COMPLETE, you can click into it and see options to download the file or explore it using the replay tool, just like with Webrecorder's browser extension.

A few of Browsertrix's fancier options:

- including scraping links one "hop" out = links on the page you scrape that go to other websites
- set limits on crawling (e.g. stop after a certain amount of time, file size, or number webpages is reached)
- create "browser profiles" that let the crawler use a login or accept certain cookies (e.g. if you wanted to crawl a site that's behind a paywall)
- give Browsertrix a list of URLs to crawl

Visit docs.browsertrix.com to learn all the possible features and settings.

*DIY web archiving is 100% more fun with collaborators of any species*

# Webrecorder quality assurance

Browsertrix's "quality assurance" (QA) feature helps you check if your archive crawl worked, without needing to review each of potentially thousands of pages individually —instead, the tool helps surface the pages most likely to need manual checking.

This QA process compares screenshots and extracted text taken from the live website during crawling against pages loaded from the archive in replay, to highlight pages most likely to have issues stemming from either the crawling process or replay. You can:

- view screenshots side-by-side of such pages & use a slider to visually check for missing images, embeds, etc.
- go to the "text" tab to compare text across the pages and see any differences highlighted with red, to be sure the issue isn't just that the screenshot was taken before the page finished loading (i.e. the crawl has the entire image correctly)
- or, the easiest approach: view the archived page in the web player to see if it looks correct

Things disappear from the internet!
Web archiving **buys you time** to capture things while they are available, then work to translate them into other media later if you need to, or put them back online if things change politically.

ILYA KREYMER, WEBRECORDER PROJECT CREATOR

# Step 2: make those copies usable in the future

Making copies from the web is a 1st step; the 2nd step is naming, recording details about, and storing that data so you (and potentially others) can locate it and remember what it is, in the future.

Metadata, sometimes described as "a love letter to the future", is data about data, e.g. details about when a webpage screenshot was captured and what it shows. Picture happening on your data in four years and not remembering anything about it—how can you title, record details so you understand when and why it was captured, and how you might want to use it?

WHAT IS THE METADATA THAT'S GOING TO BE MOST RELEVANT TO YOUR COMMUNITY OR TO YOUR USERS OR TO YOUR AUDIENCE?

ANNA KIJAS, SUCHO* CO-FOUNDER (W/QUINN DOMBROWSKI, SEBASTIAN MAJSTOROVIC); LIBRARIAN AT TUFTS UNIVERSITY

Consider:

- SCOPE: How exhaustive is what your data represents? (e.g. entire website, or just several of site's pages)
- CONTEXT: If you found the data using a search, what search terms did you use? If using a big engine like Google, include context that can impact what results you were shown (e.g. date, location, using a VPN...)
- Is there any chance someone other than you will get to access these (i.e. need to understand what they are)?

** SUCHO = Saving Ukrainian Cultural Heritage Online Initiative, sucho.org*

# Best practices for file naming

It's important to be able to tell from their filenames what the difference is among your files—if you leave things with default filenames like "screenshot1.png" now, you're putting off labor to the future, when you'll need to both open files and be lucky enough to remember what they represent, to understand what they are. Decide a filenaming convention BEFORE you start archiving!*

What folder, subfolders, folder names will you create?

- Will you be archiving the same website, or portions of the same website, repeatedly in the future? You may want different folders for each date you archive.
- Consider ease of using that archive: are you capturing material on topics from across many sites, or a big social media platform? You may want a folder per topic, platform (e.g. Mastodon), and/or site type (e.g. "newspaper websites").

Filenaming conventions include:

- Capture date (use YYYY-MM-DD for easy date sorting)
- Institution or website name (so you can find site again)
- Domain suffix (.edu, .org, .com etc.)

For example, naming a WACZ file of https://amhersthistory.org:

- UNHELPFUL filename: webarchive.wacz
- HELPFUL filename: 2024-11-24-amherst-history-org

** To understand what/why of metadata description practices in depth, see Dooley & Bowers' resource at doi.org/10.25333/C3005C*

# Best practices for describing data

Consider what level of description future you/others may need to understand what a capture is.

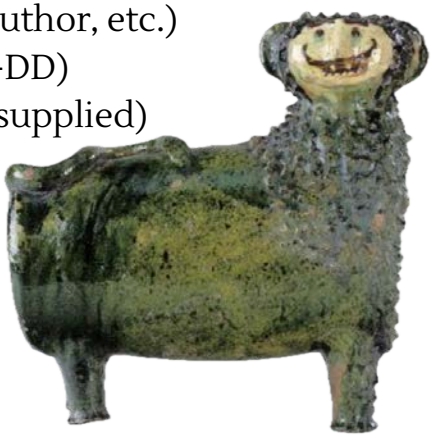
*If you're using Webrecorder, the tool may autogenerate some of this metadata!*

For captures of entire sites:

- Title of site
- Date of capture (YYYY-MM-DD)
- Website creator/owner
- Host institution (if you find the data in a particular university or other group's archive); host's location
- Extent/Size (i.e. MB, TB)
- Rights (e.g. copyright, license)
- Name of archiver
- Site URL

For captures of individual items (e.g. a photo, PDF...):

- Title of object
- Creator of object (i.e. artist, author, etc.)
- Date of creation (YYYY-MM-DD)
- Subject heading (original or supplied)
- Original description (if any)
- Source URL
- Object filename
- Relation (e.g. is Part of series or collection, has Part)



See this example from SUCHO's* "Exploring Ukrainian Cultural Heritage Online": gallery.sucho.org/items/show/25

** SUCHO = Saving Ukrainian Cultural Heritage Online Initiative, sucho.org*

# Why consider alternatives to Big Tech tools?

TODAY IS NOT NECESSARILY THE DAY TO GO UPEND ALL OF YOUR DIGITAL INFRASTRUCTURE. BUT DEPENDING ON HOW THE WINDS SHIFT, IT'S JUST GOOD TO KNOW THAT THESE RESOURCES ARE OUT THERE, SO YOU'RE NOT LEFT SEARCHING FOR THESE THINGS IF THEY BECOME URGENT.

QUINN DOMBROWSKI, SUCHO* CO-FOUNDER (W/KIJAS & MAJSTOROVIC); LIBRARIAN AT STANFORD

We don't know yet what the legal and tech landscape will look like in the near future, for the digital tools that we've come to depend upon. We may reach a point where we need to disengage from big tech. Given the political alignments of the big tech companies, there's reason to think they may not be all that invested in preserving user privacy, etc., including if they receive requests from the government.

For SUCHO*, we used Google Drive as our major storage and collaboration tool; we used Google Slides for the workshop this zine is based on. You may reach a point where you need to not use Google products, to return to a sense of security and ownership of your data.

Here are a few alternatives we've used and recommend...

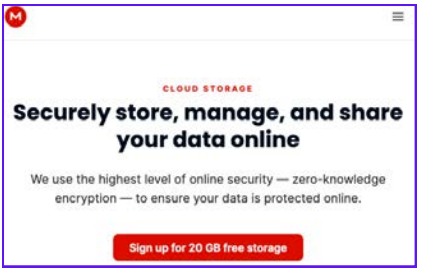
** SUCHO = Saving Ukrainian Cultural Heritage Online Initiative, sucho.org*

# Alternatives to Big Tech tools

Collaborate: Cryptpad (cryptpad.digitalcourage.de)

- Encrypted & open source
- Run it on your own server, or use a version hosted out of Germany (i.e. not U.S. servers)
- Allows basically same stuff Google Drive does (e.g. spreadsheets, collaborative doc editing)


Store: Mega.io (mega.io/storage) Longstanding, well-known storage option w/encrypted data. Even if government demands handing your data over, user-based encryption limits what can be shared. Public transparency report on how many takedown notices they've received & how many granted; history of resisting user-data government requests.




- 20GB storage is free; pay for more

Email: We currently do not have a recommendation for email, due to the embrace of fascism by the CEO of our previous email platform recommendation. If you have suggestions for email options, please share with us? (Not ~~Proton Mail.)~~

# Considerations for safety & privacy

Who could be harmed if your data is accessed, crawled, ingested into a database, AI training set, or other tool?

- How might data that preserves work by or connects at-risk groups (e.g. trans folks) be used to track, target those groups if accessed by the wrong people?
- Even for data already theoretically findable online, gathering may amplify it to more folks than were previously aware of it, or make targeting folks easier.

It's important to archive these things! But think about where/how you store them, who you share them with, and if there are ways to reduce risk.

Personal safety options for when archiving:

- Use a VPN, if you can. Some free options may be great, but some might be paid by their uses of your data or less secure.
- Using dedicated logins for web archiving can prevent your personal account from getting limited or even banned and can help provide more anonymity
- Work for a government-funded group (including public universities)? Anyone can make a FOIA request accessing Slack chats, email, Zoom transcripts, etc. that aren't protected by a limited set of student-privacy and HR rules. Do non-work DIY web archiving firmly outside work resources if you can.

# Further resources

1. Event slides: tinyurl.com/diy-web-archiving-11-2024
2. Tools:
    a. Webrecorder. webrecorder.net
    b. Mega.io storage: mega.io/storage
    c. Protonmail: proton.me/mail
    d. Cryptpad: cryptpad.digitalcourage.de
3. Tutorials
    a. SUCHO's Tutorials wiki.sucho.org/en/tutorials
4. Example: SUCHO "Metadata template & data dictionary" tinyurl.com/sucho-metadata-template
5. "Awesome Web Archiving." github.com/iipc/awesome-web-archiving
6. Dooley & Bowers "Descriptive Metadata for Web Archiving: Recommendations" doi.org/10.25333/C3005C
7. End of Term Archive: eotarchive.org/contribute
8. IIPC. "Web Archiving." netpreserve.org/web-archiving
9. Free zines on activist digital cultural heritage archiving, safety, surveillance: zinebakery.com/subsets/activist-digital

# “DIY Web Archiving: For the Web (& World) You Care About”

A friendly tutorial by Tessa Walsh, Ilya Kreymer (Webrecorder); Quinn Dombrowski, Anna Kijas (SUCHO); & Amanda Wyatt Visconti.

Version 4. January 30, 2025.
Charlottesville, VA, U.S.A.
Zine Bakery Bakeshop (Collab) Zines #2
zinebakery.com/homemade-zines/bakeshop-2-diywebarchiving

Based on the 11/25/2024 virtual workshop co-sponsored by WebRecorder, SUCHO, and ACH; taught by Walsh, Kreymer, Dombrowski, & Kijas; with additional text/resources from & zineified by Amanda Wyatt Visconti.

Thanks @ co-sponsors of the workshop this zine is based on:

- SUCHO (Saving Ukrainian Cultural Heritage Online), sucho.org
- Webrecorder, webrecorder.net
- ACH (Association for Computers & the Humanities), ach.org

A CC BY-NC* licensed zine:

## Please DO reprint, remix, redistribute—and archive :) this zine freely!

No requirement to ask us first, though we do absolutely love hearing you're sharing or using this zine.

CITE THIS ZINE

Dombrowski, Quinn, Tessa Walsh, Anna Kijas, Ilya Kreymer, Amanda Wyatt Visconti. "DIY Web Archiving: For the Web (& World) You Care About". Zine Bakery Bakeshop #2; Version 4; January 30, 2025. Charlottesville, VA, U.S.A. https://zinebakery.com/homemade-zines/bakeshop-2-diywebarchiving

ON SOCIAL MEDIA

(Amanda made the zineification, so is the best person to contact re:zine typos, accessibility improvements, etc.)

Quinn Dombrowski: @quinnanya.me
Tessa Walsh: @bitarchivist.net
Anna Kijas: @akijas.bsky.social
Ilya Kremer: @ilya.webrecorder.net
Amanda Wyatt Visconti: literaturegeek.bsky.social

* CC BY-NC is a Creative Commons license stating you can share and remix this zine freely, as long as you both credit the authors + do not ask for money for the zine (or bundles including it)